

# Northumbria Research Link

Citation: Zhao, Jiaojiao, Han, Jungong and Shao, Ling (2018) Unconstrained Face Recognition Using a Set-to-Set Distance Measure on Deep Learned Features. IEEE Transactions on Circuits and Systems for Video Technology, 28 (10). pp. 2679-2689. ISSN 1051-8215

Published by: IEEE

URL: <https://doi.org/10.1109/tcsvt.2017.2710120>  
<<https://doi.org/10.1109/tcsvt.2017.2710120>>

This version was downloaded from Northumbria Research Link:  
<http://nrl.northumbria.ac.uk/id/eprint/39186/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)

# Unconstrained Face Recognition Using A Set-to-Set Distance Measure

Jiaojiao Zhao, Jungong Han, and Ling Shao, Senior Member IEEE

**Abstract**—Recently considerable efforts have been dedicated to unconstrained face recognition, which requires to identify faces “in the wild” for a set of images and/or video frames captured without human intervention. Unlike traditional face recognition that compares one-to-one media (either a single image or a video frame) only, we encounter a problem of matching sets with heterogeneous contents containing both images and videos. In this paper, we propose a novel Set-to-Set (S2S) distance measure to calculate the similarity between two sets with the aim to improve the recognition accuracy for faces with real-world challenges such as extreme poses or severe illumination conditions. Our S2S distance adopts the  $kNN$ -average pooling for the similarity scores computed on all the media in two sets, making the identification far less susceptible to the poor representations (outliers) than traditional feature-average pooling and score-average pooling. Furthermore, we show that various metrics can be embedded into our S2S distance framework, including both predefined and learned ones. This allows to choose the appropriate metric depending on the recognition task in order to achieve the best results. To evaluate the proposed S2S distance, we conduct extensive experiments on the challenging set-based IJB-A face dataset, which demonstrate that our algorithm achieves the state-of-the-art results and is clearly superior to the baselines including several deep learning based face recognition algorithms.

**Index Terms**—Face recognition, IJB-A, S2S Distance,  $kNN$ -average Pooling.

## I. INTRODUCTION

Recent years have witnessed an explosion of face media available on the Internet. Picasa photo albums and Facebook, for example, create thousands of face images/videos every day, most of which are captured without control of age, pose, illumination, occlusion and expression [1], [2], [3], [4]. This high volume of real-world face images and videos now requires face recognition, more than ever, to handle large quantities of faces and meanwhile remain sufficiently accurate even when provided with images/videos taken under unconstrained conditions.

There have been significant breakthroughs on applications and techniques of face recognition under unconstrained environments over the past few years, which are also in accordance with the progress of the face datasets. At the first phase of unconstrained face recognition, a single image setting is

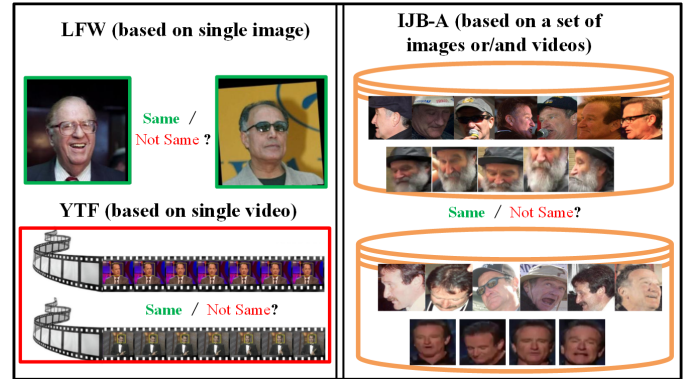


Fig. 1: Difference between IJB-A dataset and previous unconstrained face datasets: LFW is based on images; YTF is based on videos; IJB-A is based on sets of images or/and videos. Each set contains full variations in face pose, expression, illumination and occlusion issues.

always used in the datasets such as the Labeled Faces in the Wild dataset (LFW) [5] (see Fig. 1 for an example) and the Public Figures (PubFig) dataset [6]. Both datasets consist of face images harvested from news websites of labeled people, which can be seen as a key step towards identifying faces in an unconstrained condition [7]. Early recognition methods dealing with this sort of datasets simply adapted the techniques available for the environments under human control to this application, which, not surprisingly, failed to obtain high accuracy. Stepping into the second phase, datasets such as the YTF [8] (see Fig. 1 for an example) over a video attracted much attention. Recently, owing to the exploration of deep learning [9], [10], the recognition accuracies on these datasets reached almost one hundred percent.

Despite the rapid progress, unconstrained face recognition is hardly considered as a solved problem because the datasets mentioned above are still far from reality. For such unconstrained environments, there remains a large gap between automatic machine recognition and human recognition. We consider the more challenging and practical unconstrained face recognition [7] on the IARPA Janus Benchmark A (IJB-A) dataset (see Fig. 1). The IJB-A dataset is a mixture of images and videos with full pose variation, complex expression and illumination. Additionally, IJB-A is a set-based dataset, different from older image-to-image or video-to-video datasets, meaning that it takes a set (heterogeneous contents containing both images and videos) as the smallest unit of representation. This kind of setting does reflect the real-world biometric scenarios, thereby attracting a lot of attention after its release. In

This research was supported by the Royal Society Newton Mobility Grant IE150997. (Corresponding authors: Jungong Han.)

Jiaojiao Zhao is with the Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK (e-mail: jiao-jiao.zhao@northumbria.ac.uk).

Jungong Han is with the School of Computing and Communications, Lancaster University, Lancaster, UK (e-mail: jungonghan77@gmail.com).

Ling Shao is with the School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K. (Email: ling.shao@ieee.org).

this case, besides the old problem of how to extract invariable and discriminative features, finding a solution to match two sets of media is also challenging. Most of the existing methods until now mainly include feature pooling [11], [12], [13] and score pooling [25], [26], [28], where the former suggests aggregating features over all images in a set while the latter aggregates the pair-wise similarity scores of two compared sets. However, neither of the two measurements performs well when the faces of certain subjects are with many extreme poses or other variations, which occur frequently in reality.

Aiming to address this Set-to-Set matching problem, in this paper, we propose a simple but effective S2S distance by leveraging the  $kNN$ -average pooling. Our recognition system starts by a deep feature extraction step making use of a VGG-16 deep network and a transfer learning technique to compensate for lack of training data. The trained model in turn is used to acquire features from all images over a set in the test phase. We employ the proposed S2S distance to calculate the similarity of two sets, and in the end make decision for the recognition tasks. The overview of the entire system is illustrated in Fig. 2 and our contributions can be summarized as the following three points:

- The primary contribution lies in a novel Set-to-Set distance employing  $k$  Nearest Neighbor ( $kNN$ )-average pooling to measure the similarity between two sets of face media. The S2S distance is simple but very effective and robust to the outliers.
- Our framework built on the S2S distance is so flexible that different metrics including both pre-defined and learned ones can be incorporated. The experimental results reveal that there is no need to learn a particular metric for such a challenging dataset wherein many subjects are with very limited media samples, even sometimes with only one sample.
- Compared with feature-pooling, score-pooling and even some deep learning methods, our simple S2S distance helps to achieve the state-of-the-art results on IJB-A which is the only set-based face dataset.

The rest of the paper is organized as follows. In Section II we review the latest works on the IJB-A dataset and make a brief discussion. Afterwards, our motivation and method are presented in Section III. Section IV details the experiments to demonstrate the effectiveness and flexibility of our approach followed by a conclusion and our future work in the last section.

## II. RELATED WORK

As described in the previous section, earlier recognition methods obtained quite low accuracies under the unconstrained settings. In order to improve the recognition performance on the datasets over images, some researchers dedicated to finding better face representations or descriptors. In [14], the authors used Fisher vectors on densely sampled SIFT and then compressed the encoding to a small representation. Apart from the descriptor-based methods, feature selection/compression and metric learning [15], [16], [17] also made some contributions to improve the face recognition under unconstrained environments.

Furthermore, some previous methods dealing with the set-based setting, e.g., the YTF benchmark, were also developed. In this database, the probe and the gallery were typically comprised of multiple frames from the same video. The simple way was to generate one feature by computing the mean of all features in each set and then compare the two aggregated features of the two given sets. One elaborate method designed for this purpose was convex hull [18] which performed well when many frames were available in sets. Under the assumption that the elements in a set may lie close to a linear subspace, subspace-based methods [19] were introduced. To overcome the drawbacks of the traditional kernel-based methods, Huang *et al.* [20] proposed a method to learn the projection metric directly from a Grassmann manifold. Moreover, various distribution based representations were considered, such as the Bag of Features [21](BF) and Vector of Locally Aggregated Descriptors (VLAD) [22].

Most of the latest evaluations on the IJB-A dataset relied on deep Convolutional Neural Network (DCNN). Chen *et al.* [11] designed a DCNN with small filters trained on CASIA-WebFace [23] and then learned a joint Bayesian metric to measure the similarity between two faces. Similar to the work using metric learning, in [12], the authors coupled a DCNN with a Triplet Distance Embedding (TDE). To tackle the Set-to-Set problem, both works applied the feature-average pooling, by which the features of all media in a set are averaged to form an overall feature representation. The authors of the later work further extended the AlexNet [24] to a DCNN through embedding a Triplet Probability (TP) [13]. Meanwhile, they replaced the feature-average pooling with the media-average pooling so as to produce a better representation of a set. A bilinear CNN (B-CNN) [25] was applied due to the fact that face recognition can also be viewed as a fine-grained classification problem. In such a framework, one-versus-rest linear SVM classifiers were trained on the gallery set for the identification task. In addition, the authors tried two kinds of max pooling methods within a set: score-max pooling and feature-max pooling. Alternatively, [26] exploited 3D rendering to generate multiple face poses from the input image and then produced multiple pose-specific features by several pose-specific DCNN models. They simply adopted a direct feature-to-feature comparison in a pair-wise fashion and used the softmax weights to fuse all the scores for a set. Hassner *et al.* [27] illustrated a PoolingFaces solution using DCNN to encode the pooled images falling into the same bin according to pose and image quality. The work reported in [28] learned the Pose-Aware models for faces with different poses using CASIA-WebFace. It enables a weighted average of the pair-wise scores for each set, where each weight is specified to be an exponential function of the score. In order to solve the problem of a small dataset, Iacopo *et al.* [29] presented a simple synthesized means to augment the training data. Moreover, a cascaded face search system [30] provided an efficient framework when dealing with large-scale face data. An end-to-end system consisting of face detection, alignment and recognition based on deep convolutional networks was proposed in [31]. In [32], a template adaption method integrated the DCNN into a set specific linear SVM, in which

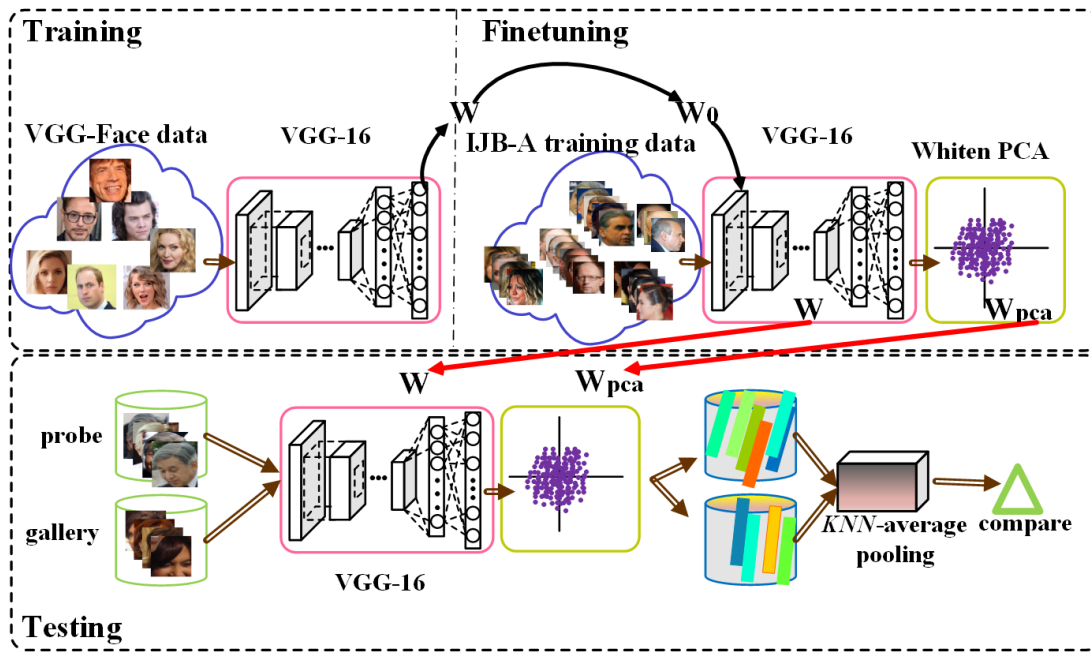


Fig. 2: Overview of the entire system with three phases are included. Training: VGG-16 trained on VGG-Face data; Finetuning: Finetune pre-trained model on IJB-A training data; Testing: Using fine-tuned model to extract features and applying  $kNN$ -average pooling on extracted features.

SVM classifiers [33] were generated for both gallery and probe sets. However, it does not seem to be practical to find appropriate negative and subject-disjoint positive samples for the probe set, because the subject labels of the probe set were not always available in real applications.

### III. MOTIVATION AND METHOD

#### A. Motivation

As described above, in order to push the unconstrained face recognition research close to the real-world applications, the IJB-A dataset over sets of media (images and/or videos) was released. It is understandable that a set of media rather than a single image/video provided for a subject means that more information can be available for recognition, which should be helpful for achieving better results. However, it is not always like this, especially in reality due to many uncontrolled factors. On the one hand, the great variety in age, pose, expression, illumination and other conditions, makes the set heterogeneous in contents. On the other hand, incorrect results caused by preprocessing such as face detection and alignment may introduce noise to sets. Even by looking at the ground-truth boxes provided by IJB-A, one can easily notice that they are unstable and some of them are labeled incorrectly. At the same time, we find a few persons are given wrong subject labels in the dataset. In such a Set-to-Set matching problem, given a noisy but practical dataset, the key is to avoid skewing matching scores caused by the complex factors. Essentially, we should decide which media are useful for comparison and how to weigh the similarity scores of different cross-set media pairs. From this point of view, previous strategies adopting either feature pooling or score pooling cannot generalize well in many practical cases. We make an in-depth analysis

following a preliminary definition. A feature representation  $z = f(x)$  is a mapping  $f(x) \in \mathbb{R}^d$  from an image or a frame  $x$  to an encoding  $z$  with dimensionality  $d$ . Let  $\tilde{z} = \frac{1}{m} \sum_x f(x)$  be the average of features of images or frames in media  $S$ , such as the average encoding for all  $m$  frames in a video.

**Feature-average Pooling** is proposed as a useful approach for endowing the features with invariant properties. Both image-based feature-average pooling and media-based feature-average pooling were used in [24], in which the former takes a component wise average of the features over all the images or/and frames in a set  $\mathbf{P} = \{z_1^P, z_2^P, \dots, z_n^P\}$ . Hence, the final representation is  $F_P = \frac{1}{\|\mathbf{P}\|} \sum_z z^P$ .  $\|\mathbf{P}\|$  is the number of images or/and frames within the set. The latter firstly conducts an intra-media average and then combines them via an inter-media average. Hence, there are  $\mathbf{P} = \{\tilde{z}_1^P, \tilde{z}_2^P, \dots, \tilde{z}_n^P\}$  and  $F_P = \frac{1}{\|\mathbf{P}\|} \sum_{\tilde{z}} \tilde{z}^P$ , where  $\|\mathbf{P}\|$  is the number of media in the set. The similarity between  $\mathbf{P}$  and  $\mathbf{G}$  can be defined as:

$$\text{sim}_{\text{feature\_aver}} = K(F_P, F_G), \quad (1)$$

Here,  $K(x, y)$  represents a kernel similarity measurement. Actually, the media-based feature-average pooling can be considered as a strategy to divide a set into several clusters and get one representation for the images with similar properties. For example, firstly taking the average of all frames in a video can help reduce the effect of noise from the video to the whole set. However, either of the two ways is not suitable for the sets with heterogeneous contents. For instance, if there are a few images with extreme poses and many with frontal faces in a set, the feature after averaging will be similar to a frontal face. In another example, suppose the dataset contains many faces with extreme poses and a few frontal faces, its average feature



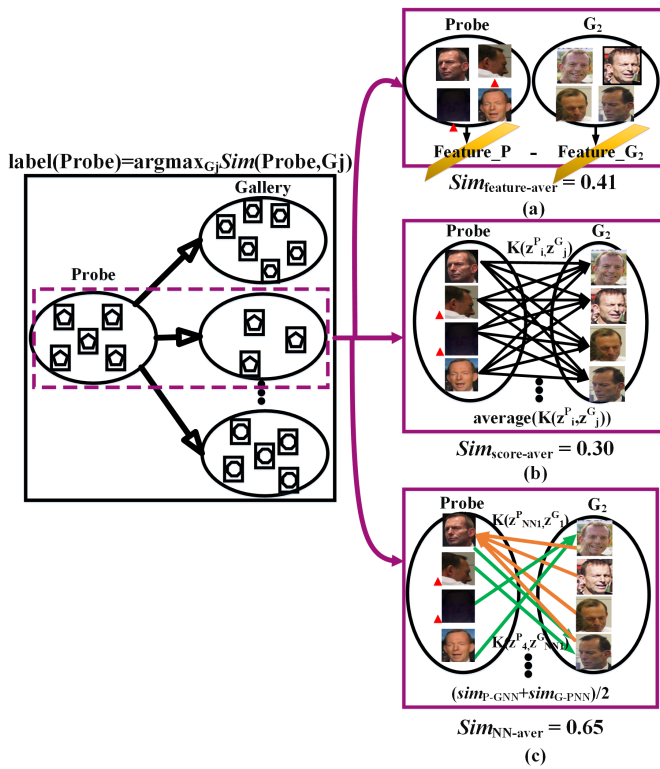


Fig. 3: Illustrations of the three pooling methods. (a) feature-average pooling; (b) score-average pooling; (c)  $kNN$ -average pooling ( $k = 1$ ). In the probe set, the faces marked by red triangle can be seen as bad examples. In (c), the green line is from each sample in the probe set to its corresponding  $NN$  in gallery set; the orange line is from each sample gallery set to its corresponding  $NN$  in the probe set.

is likely to be close to an extreme pose. For both cases, the similarity score may not be high if any of two average features is adopted. Thus, the feature-average pooling may cause the loss of key information.

**Score-average Pooling** means the final feature is generated by averaging the calculated pairwise similarity scores (based on media features) of two sets. Given a probe set  $P = \{\tilde{z}_1^P, \tilde{z}_2^P, \dots, \tilde{z}_n^P\}$  and a gallery set  $G = \{\tilde{z}_1^G, \tilde{z}_2^G, \dots, \tilde{z}_l^G\}$ , the similarity score is represented as:

$$sim_{score-aver} = \frac{1}{\|P\|} \cdot \frac{1}{\|G\|} \sum_{i,j} K(\tilde{z}_i^P, \tilde{z}_j^G), \quad (2)$$

Obviously, the operation is fragile to outliers because it assigns equal weights even for outliers, which will definitely affect the final decision. Moreover, score-max pooling is tried in [25] but it is not robust to noises either.

### B. $kNN$ -average pooling

In view of the above analysis, we can draw a conclusion that either of feature-average pooling and score-average pooling may not properly handle the case when two sets have variable contents and noises. Aiming to solve this problem, we propose a simple but effective method called  $kNN$ -average pooling.

TABLE I: Pairwise Cosine Similarity Scores Between Probe Set and Gallery Set. The red boxes label the low scores from extreme poses and noises. Finally, our  $kNN$ -average pooling uses the scores in the green circles

Gallery \ Probe				
Probe 1	0.6427	0.4939	0.3524	0.6336
Probe 2	0.1289	0.1365	0.1885	0.3859
Probe 3	0.2032	0.1508	0.1663	0.1108
Probe 4	0.5208	0.4306	0.0567	0.2035

**$kNN$ -average Pooling** basically takes advantage of two kinds of pooling of the pairwise similarity scores between a probe set  $P = \{\tilde{z}_1^P, \tilde{z}_2^P, \dots, \tilde{z}_n^P\}$  and a gallery set  $G = \{\tilde{z}_1^G, \tilde{z}_2^G, \dots, \tilde{z}_l^G\}$ . For each media item  $\tilde{z}_i^P$  in the probe set, we first calculate the similarity between  $\tilde{z}_i^P$  and its  $k$  nearest neighbors  $\tilde{z}_{NN_k}^G$  in the gallery set and sum them:

$$sim_{\tilde{z}_i^P} = \sum_{j=1}^k K(\tilde{z}_i^P, \tilde{z}_{NN_j}^G), \quad (3)$$

and then take the average of all the scores as the similarity from the probe to the gallery

$$sim_{P-GNN} = \frac{1}{\|P\|} \sum_i sim_{\tilde{z}_i^P},$$

Likewise, we do the same for the gallery set to get the similarity from the gallery to the probe

$$sim_{\tilde{z}_j^G} = \sum_{i=1}^k K(\tilde{z}_{NN_i}^P, \tilde{z}_j^G),$$

$$sim_{G-PNN} = \frac{1}{\|G\|} \sum_j sim_{\tilde{z}_j^G}.$$

Finally, the similarity between the two sets is

$$sim_{kNN-aver} = (sim_{P-GNN} + sim_{G-PNN})/2, \quad (4)$$

Generally, feature-average pooling and score-average pooling can be seen as comparing the distributions of two sets. However, it is likely to get a low similarity score in the case that the compared two sets do not contain lots of media or they include severe noises. Such a phenomenon can be illustrated in Fig. 3, where the probe set and the gallery set belong to the same subject. In the probe set, there exists an extreme pose and a noisy image caused by poor illumination. We use

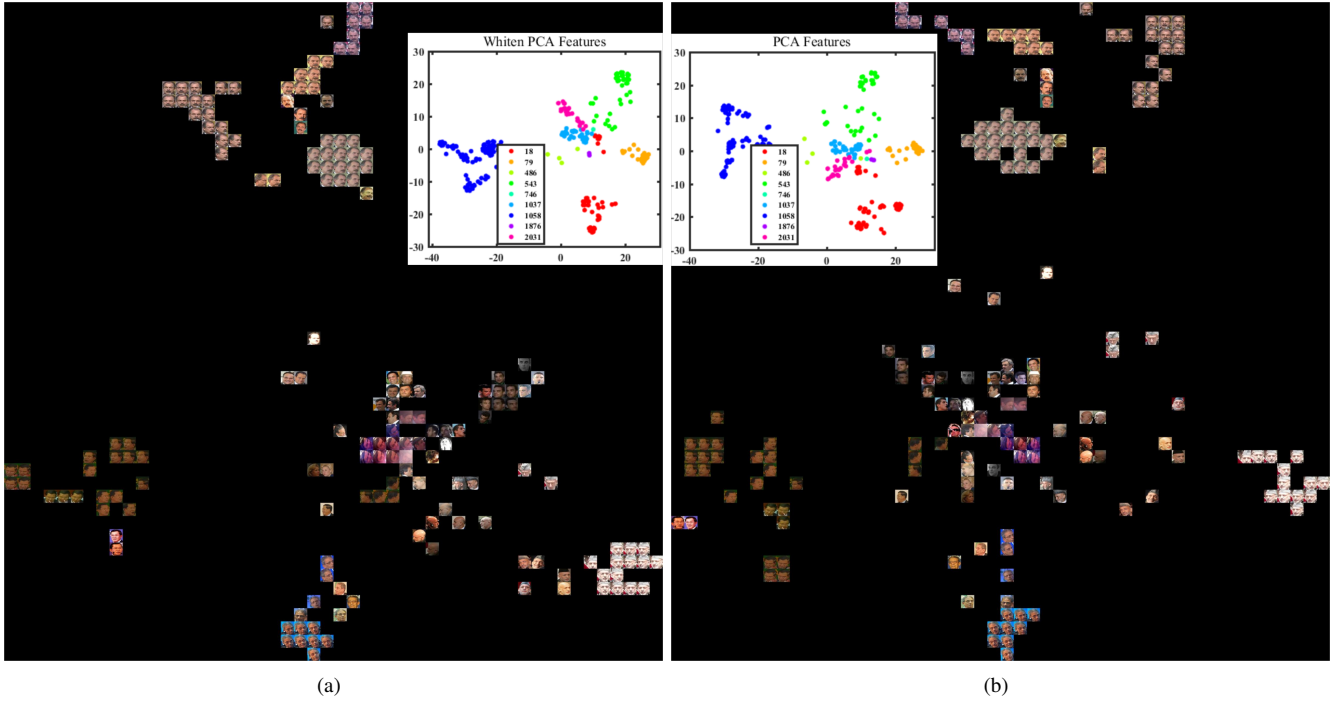


Fig. 4: t-SNE visualization of deep features of 10 classes randomly selected from test dataset of split 1 after whiten PCA (a) and PCA (b).

deep features (from a VGG-face model fine-tuned on IJB-A training data) and Cosine similarity metric to calculate the three similarity measurements. Here, we set  $k = 1$ . From the figure, both feature-average pooling and score-average pooling obtain lower similarity scores for the two sets from the same subject. On the contrary,  $kNN$ -average pooling gets a higher matching score. Intuitively,  $kNN$ -average pooling behaves like a weighting strategy to select more good-samples and less bad-samples. We also list the pairwise scores based on Cosine metric in TABLE I. It can be observed that the faces with the extreme poses and the noisy images always generate low scores. Next, we present our idea from a naive probabilistic formulation perspective.

### C. Naive probabilistic formulation

As we mentioned above, face recognition can be regarded as a classification problem, which checks whether the probe is the same subject class with a reference or which subject class in the gallery the probe belongs to. Given a new probe image set  $\mathbf{P}$ , we need to find its subject class  $\mathbf{G}$ . Based on probability theory, the maximum-a-posteriori (MAP) classifier minimizes the average classification error [34], [35]:

$$\hat{\mathbf{G}} = \arg \max_{\mathbf{G}} p(\mathbf{G}|\mathbf{P}).$$

Usually, Maximum-Likelihood (ML) is used to represent MAP when the class prior  $p(\mathbf{G})$  is uniform:

$$\hat{\mathbf{G}} = \arg \max_{\mathbf{G}} p(\mathbf{G}|\mathbf{P}) = \arg \max_{\mathbf{G}} p(\mathbf{P}|\mathbf{G}).$$

If a probe set  $\mathbf{P}$  contains some elements  $\tilde{z}_1^P, \tilde{z}_2^P, \dots, \tilde{z}_n^P$ , under the Naive-Bayes assumption, the probability of  $\mathbf{P}$  belonging to  $\mathbf{G}$  is modeled as:

$$p(\mathbf{P}|\mathbf{G}) = p(\tilde{z}_1^P, \tilde{z}_2^P, \dots, \tilde{z}_n^P|\mathbf{G}) = \prod_i p(\tilde{z}_i^P|\mathbf{G}).$$

Combine the two formulas and take the log format,

$$\hat{\mathbf{G}} = \arg \max_{\mathbf{G}} \log(p(\mathbf{G}|\mathbf{P})) = \arg \max_{\mathbf{G}} \sum_i^n \log p(\tilde{z}_i^P|\mathbf{G}) \quad (5)$$

According to Eq. 5, we need to calculate the probability density  $p(\tilde{z}^P|\mathbf{G})$  of elements  $\tilde{z}^P$  in gallery  $\mathbf{G}$ . Referring to [34], we get an estimation  $\hat{p}(\tilde{z}^P|\mathbf{G})$  as:

$$\hat{p}(\tilde{z}^P|\mathbf{G}) = \frac{1}{l} \sum_{j=1}^l K(\tilde{z}^P, \tilde{z}_j^G) \quad (6)$$

Furthermore, an accurate approximation of Eq. 6 using (few)  $k$  largest elements in the sum is given below:

$$p_{NN}(\tilde{z}^P|\mathbf{G}) = \frac{1}{l} \sum_{j=1}^k K(\tilde{z}^P, \tilde{z}_{NN_j}^G) \quad (7)$$

These  $k$  largest elements correspond to the  $k$  nearest neighbors of an element  $\tilde{z}^P \in \mathbf{P}$  within the elements  $\tilde{z}_1^G, \tilde{z}_2^G, \dots, \tilde{z}_l^G \in \mathbf{G}$ . This is the probability model of our  $kNN$ -average pooling method. Due to that  $l$  is a constant, we approximate a similarity defined in Eq. 3. To match two sets, we apply a symmetric format in Eq. 4.

### D. Pre-defined metrics embedded vs Learned metrics embedded

As described in Fig. 2, the next step after defining the format of Set-to-Set distance is to find an appropriate metric measuring the distance. Our framework based on the proposed

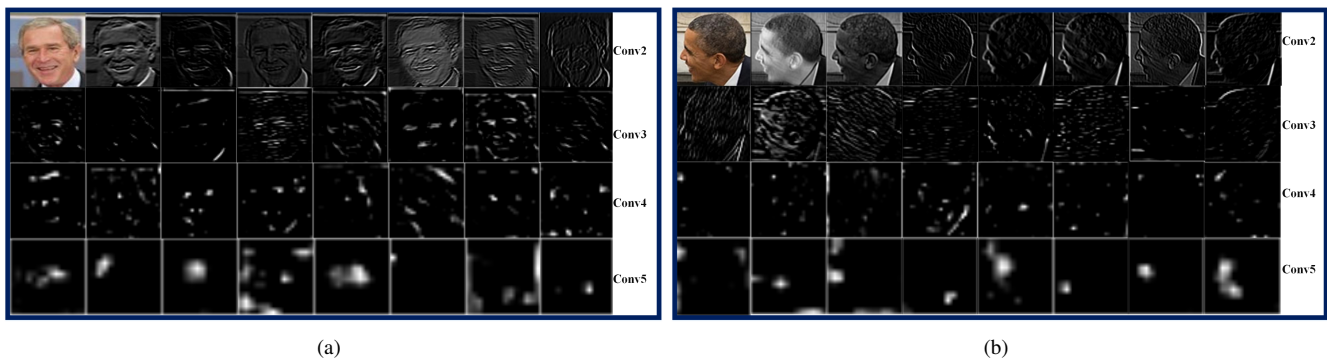


Fig. 5: Visualization of deep features from Conv2, Conv3, Conv4 and Conv5. (a) Frontal face; (b) Profile face. Most of the reaction hits on eyes, noses, mouths and ears which are most important for face recognition.

$kNN$ -average pooling is very generic and flexible, which allows switching between several metrics (both pre-defined and learned).

Generally, supervised metric learning considers a classification problem, whose objective is minimizing the distance between intra-class samples and maximizing the distance between inter-class samples. However, there are many subjects with only one sample in the IJB-A dataset. In this case, metric learning does not seem to present a satisfactory performance due to lack of training data. Fortunately, our S2S distance embedded with the unsupervised metric can achieve fabulous results. In this paper, we endow the kernel function  $K(x, y)$  three different metrics including Euclidean distance [36], Cosine similarity [37] and Joint Bayesian [38], [39]. Euclidean distance and Cosine similarity are unsupervised metrics. The formulas respectively are:

$$d_{Eu}(\tilde{z}_i^P, \tilde{z}_j^G) = \|\tilde{z}_i^P - \tilde{z}_j^G\|_2, \quad (8)$$

$$d_{Cos}(\tilde{z}_i^P, \tilde{z}_j^G) = \cos(\tilde{z}_i^P, \tilde{z}_j^G) = \frac{\tilde{z}_i^P \cdot \tilde{z}_j^G}{\|\tilde{z}_i^P\| \|\tilde{z}_j^G\|} \quad (9)$$

Moreover, the Joint Bayesian metric was introduced in previous works [39], where the appearance of a face can be modeled by two parts: identity and intra-personal variation. We use LDA to approximately optimize the loss function for  $JB$  metric.

#### IV. EXPERIMENTAL RESULTS

In this section, we present details of the whole system and evaluate it on the IJB-A dataset. We plot the bar figures to show the specific results for different settings. The Receiver Operating Characteristic (ROC) curves for 1:1 face verification, the Cumulative Match Characteristic (CMC) curves for 1:N closed-set face identification and the Decision Error Tradeoff (DET) curves for 1:N open-set identification are used to show all the comparisons. Our S2S distance based on  $kNN$ -average pooling really helps to improve the results significantly. Moreover, compared to the existing methods, our system achieves the state-of-the-art results.

##### A. Implementation details

Deep features prove to be powerful in face recognition [1], [2], [3], [40], [41], [42], [43]. Due to the fact that each split of IJB-A does not contain many images, we use transfer learning to get deep representations. Specifically, we apply VGG-16 [44] deep network trained from scratch on the VGG-Face dataset [1] without overlapping with the IJB-A dataset and then fine-tune it on training dataset of each split of IJB-A. We crop the faces using the ground-truth boxes provided with IJB-A and then re-scale them to  $224 \times 224 \times 3$ . Face alignment is not done here because not all images are provided with three key points. Another reason is that it is difficult to align different profile faces and the incorrect alignment results will bring extra noises. We input the face patches of the training set into our VGG-16 network and set the base learning rate 0.001 and learning rate policy “step”. We take the penultimate layer output as the feature encoding of 4096 dimensions. Afterwards, whitened PCA [45] is used to reduce the dimension to 256. We randomly select 10 classes from the test dataset of split 1 and use t-SNE [46] to visualize its distributions after a PCA and a whitened PCA respectively (see in Fig. 4). It is clear that whitened PCA makes the intra-class distribution like a spherical shape and increases the cohesion within the same class. The final features used for  $kNN$ -average pooling are based on media sample. In our experiments,  $k = 2$  is used for all settings. If a set only has one media, we adjust  $k = 1$ .

##### B. IJB-A evaluation

IJB-A is the only public dataset over sets, which contains in total 5,712 images and 2,085 videos of 500 subjects, with an average of 11.4 images and 4.2 videos per subject [7]. They are randomly divided into 10 splits with overlap at the subject level. There are 333 subjects randomly sampled and placed in training for each split, the other 167 subjects placed in testing. There are three protocols defined on it, which are for verification, closed-set identification and open-set identification respectively. All the results reported below are averages over the 10 splits.

1) *Evaluation for face verification*: The 1:1 compare protocol is used for face verification. For each split, the number

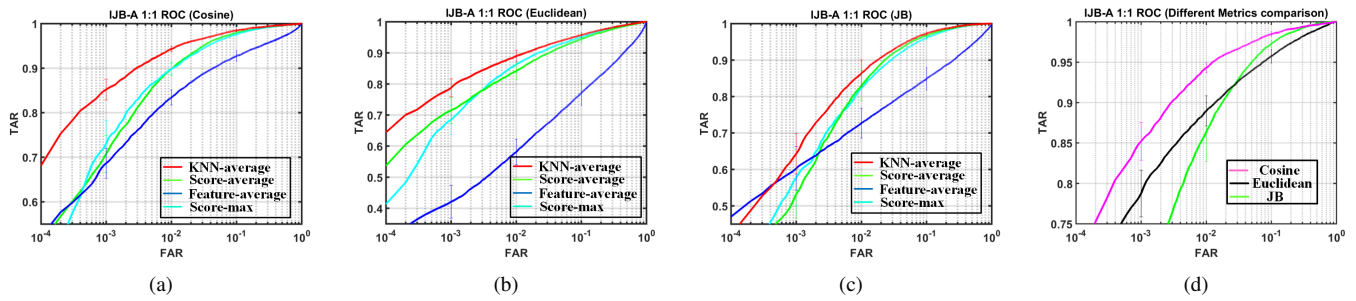


Fig. 6: IJB-A 1:1 ROC curve based on different metrics:(a) Cosine Similarity; (b) Euclidean Distance; (c) *JB* metric; (d) Comparisons between the results of *kNN*-average method based on the three metrics.

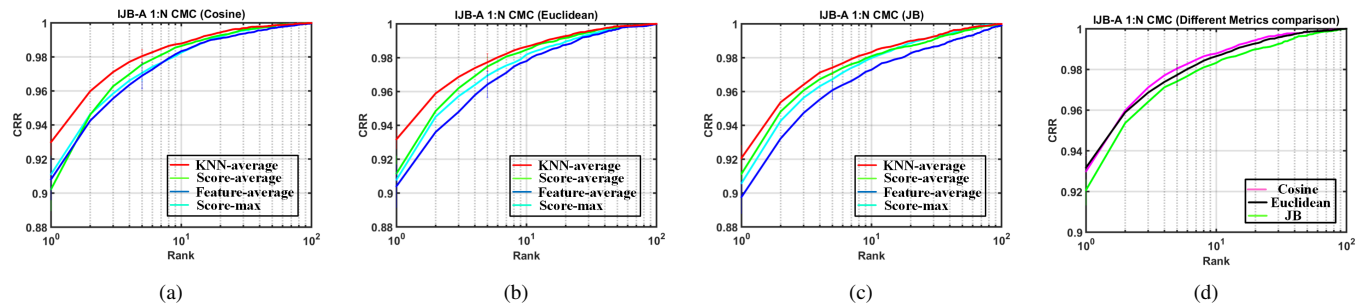


Fig. 7: IJB-A 1:N CMC curve based on different metrics:(a) Cosine Similarity; (b) Euclidean Distance; (c) *JB* metric; (d) Comparisons between the results of *kNN*-average method based on the three metrics.

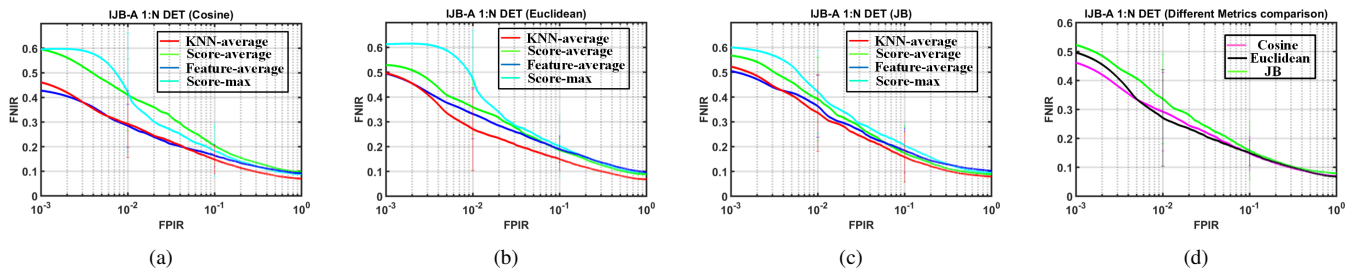


Fig. 8: IJB-A 1:N DET curve based on different metrics:(a) Cosine Similarity; (b) Euclidean Distance; (c) *JB* metric; (d) Comparisons between the results of *kNN*-average method based on the three metrics.

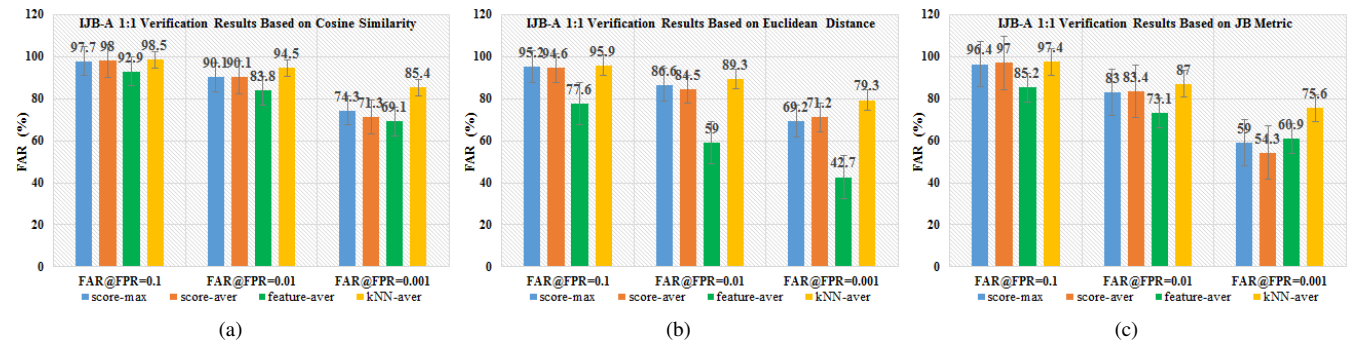


Fig. 9: IJB-A 1:1 Verification Results based on different metrics: (a) Cosine Similarity; (b) Euclidean Distance; (c) *JB* Metric.

of genuine pairs equals the number of probe sets, and a single gallery set is attached to each subject. The ROC curve is classically used to measure the verification accuracy.

Given a threshold (the independent variable), ROC draws the relationships between the true accept rate (TAR), which is the fraction of genuine pairs correctly exceeding the threshold, and



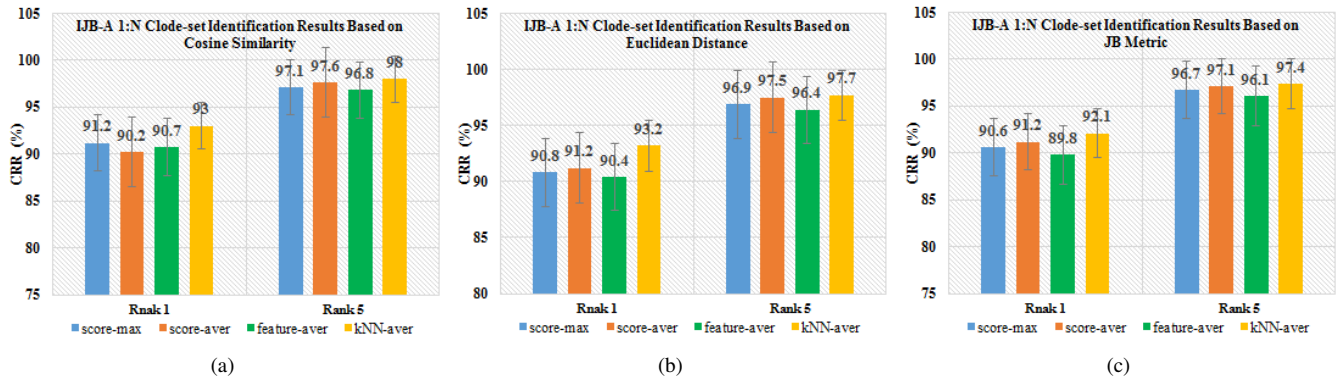


Fig. 10: IJB-A 1:N Closed-set Identification Results based on different metrics: (a) Cosine Similarity; (b) Euclidean Distance; (c) *JB* Metric.

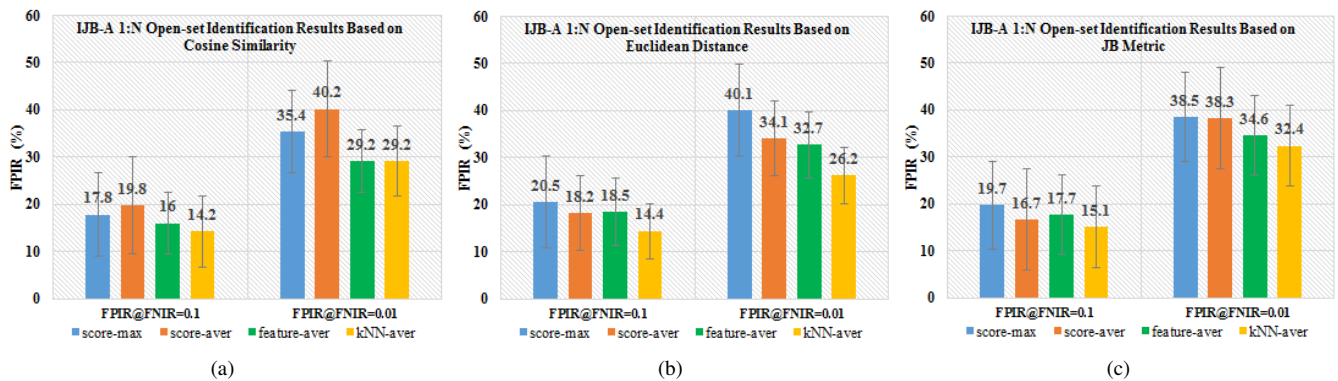


Fig. 11: IJB-A 1:N Open-set Identification Results based on different metrics: (a) Cosine Similarity; (b) Euclidean Distance; (c) *JB* Metric.

the false accept rate (FAR), which is the fraction of impostor pairs incorrectly exceeding the threshold. Using ROC curves, the following accuracies are reported: the TAR @ FAR = 0.1, 0.01 and 0.001. The higher the values are, the better the performance is.

We compare four pooling methods including score-max [25], score-average [26], [28], feature-average [11], [12], [13] and *kNN*-average pooling based on three different metrics. ROC curves are drawn for each setting in Fig. 6. Seen from the curves, our *kNN*-average pooling is clearly superior, compared to the other three pooling methods, regardless of which metric is used. It is interesting to observe that the unsupervised metric Cosine similarity is even better than the supervised metric *JB* metric. The results may reveal two important points: 1) The features learned from deep Convolutional Neural Network indeed generate good representations for most faces, except for the faces with severe noise. We show the via feature maps from Conv2, Conv3, Conv4 and Conv5 layers of one frontal face and one profile face in Fig. 5. The networks have high reaction on eyes, noses, mouths and ears which are most important for face recognition. 2) Supervised *JB* metric does not seem to work well for the case where certain subjects only include one media, because there will be not enough samples from the same class to learn the appropriate metric parameters. In contrast, the pre-defined metrics like Cosine similarity, have

no such a limitation. That's why Cosine similarity can get better results than the metric learning in this particular case. We also draw bar figures to show specific results of face verification (see in Fig. 9(a)-(c)). As observed from the bar figures, feature-average pooling causes the worst performance for most of the indexes. For the sets including many faces with big variation, this aggregation may lead to the leak of key information. Score-average and score-max obtain very similar results, however, both of which are sensitive to outliers. *kNN*-average pooling obviously outperforms them due to that it can effectively decrease the effect of noise faces. Especially for TAR@FAR=0.001, *kNN*-average has more than 10% improvement compared with the others. For face verification, Cosine similarity gives the best results using *kNN*-average pooling. For TAR @ FAR = 0.1 and TAR @ FAR = 0.01, the top accuracies are 98.5% and 94.5%, respectively. *JB* metric dose not learn a good measurement due to the fact that many subjects have only one sample in the IJB-A dataset. It actually behaves similarly with Euclidean distance in terms of the performance. Overall, our *kNN*-average pooling improves the results significantly.

2) *Evaluation for face identification*: The 1: N search protocol is used to measure the accuracy of closed-set and open-set identification. In each split, 55 randomly selected subjects appearing in the probe set are not enrolled in the gallery



TABLE II: Performance Comparison of Different Methods

Method	IJB-A 1:1 Compare			IJB-A 1:N Search			
	TAR @ FAR=0.10 (%)	TAR @ FAR=0.01 (%)	TAR @ FAR=0.001 (%)	FNIR @ FPIR=0.10 (%)	FNIR @ FPIR=0.01 (%)	Rank1 (%)	Rank5 (%)
OpenBR [7]	43.3	23.6	10.4	85.1	93.4	24.6	37.5
GOTS [7]	62.7	40.6	19.8	76.5	95.3	44.3	59.5
B-CNN [25]	-	-	-	65.9	85.7	55.8	79.6
[30]	89.5	73.3	51.4	38.7	61.7	82.0	92.9
PAMs [28]	-	82.6	65.2	-	-	84.0	92.5
DMPR [26]	91.1	78.7	-	-	-	84.6	92.7
Pooling Faces [27]	81.9	63.1	-	-	-	84.6	93.3
[12]	94.5	79.0	59.0	-	-	88.0	95.0
[42]	96.3	83.1	-	-	-	89.9	97.0
DCNNs [11]	96.7	83.8	-	-	-	90.3	96.5
[29]	-	88.6	72.5	-	-	90.6	96.2
VGG-16 [44]	-	80.5	-	33	53.9	91.3	-
$kNN$ -aver <sub>JB</sub>	97.4	87.0	65.6	15.1	32.4	89.8	96.1
$kNN$ -aver <sub>Euclidean</sub>	95.9	89.3	79.3	<b>14.4</b>	<b>26.2</b>	<b>93.2</b>	<b>97.7</b>
$kNN$ -aver <sub>Cosine</sub>	<b>98.5</b>	<b>94.5</b>	<b>85.4</b>	<b>14.2</b>	<b>29.2</b>	<b>93.0</b>	<b>98.0</b>

set. But each probe set is to be searched against the gallery sets. Generally, closed-set is for user driven searches (e.g., forensic identification). The cumulative match characteristic (CMC) is an information retrieval metric that captures the recall of a specific probe identity within top-k most similar candidates in gallery. We report rank-1 and rank-5 correct retrieval rates (CRR) for the closed setting. However, for other applications like de-duplication, some searches cannot manually be examined. There exists a tradeoff between false alarms and misses. In this case, the decision error tradeoff (DET) is plotted to measure the false positive identification rate (FPIR) and false negative identification rate (FNIR). FPIR means the false alarm rate which measures what fraction of comparisons between probe sets and non-match gallery set result in a similarity score above a given threshold  $t$ . FNIR is the miss rate measuring what fraction of probe searches fail to match a mated gallery set above  $t$ . In this paper, FPIR @ FNIR = 0.1 and 0.01 will be reported. The lower the values are, the better the performance is.

For 1: N closed-set face identification, we also compare the four pooling methods using Cosine similarity, Euclidean distance and  $JB$ . The CMC curves for all the settings can be found in Fig. 7. The  $kNN$ -average pooling presents a significant increment especially for rank-1 of the three embedded metrics. In Fig. 7(d), the comparisons for the three metrics based on  $kNN$ -average pooling are shown, in which both Cosine embedded and Euclidean embedded perform similarly but are much better than  $JB$  embedded. Fig. 10(a)-(c) display the numeric results for closed-set. Generally, the results of score-max and score-average are more or less the same. Using Cosine similarity,  $kNN$ -average pooling achieves a top accuracy of 98% for rank-5 and gets the best result of 93.2% for rank-1 if applying Euclidean distance.  $JB$  embedded gives 92.1% and 97.4% respectively for rank-1 and rank-5 based on  $kNN$ -average pooling, which help to increase the correct retrieval rate among different pooling methods.

The more challenging task is 1: N open-set identification, which requires rejecting the subjects without the enrollment in gallery. Fig. 11(a)-(c) list the results based on the four pooling

methods respectively using Cosine similarity embedded, Euclidean distance embedded and  $JB$  embedded. It is interesting to find that feature-average pooling performs better than the other two score-related pooling methods at FNIR@FPIR=0.01. It seems that feature-average representation for the subjects not enrolled in gallery can get a low score, which may help to decrease the miss rate. However,  $kNN$ -average pooling still is the best strategy for open-set. It decreases the miss rate by 1.8% for FNIR@FPIR=0.1 when Cosine similarity is embedded. Furthermore, for FNIR@FPIR=0.1 and FNIR@FPIR=0.01, there are more than 4% and 6% misses decreased, compared with the other three pooling methods using Euclidean distance. In Fig. 8, 1: N DET curves show the visual results. The results of feature-average pooling are close to those of score-average pooling but still worse than the  $kNN$ -average pooling. It reaches the same conclusion that unsupervised metrics can outperform supervised metric learning, which further supports our simple but efficient S2S distance.

3) *Failure case analysis*: We also display some failure cases based on  $kNN$ -average pooling using Cosine similarity for face verification on split 1. In Fig. 12(a), the top-30 worst matched pairs are listed. It means the two persons belonging to the same subject are verified incorrectly. In the first line, it can be seen there is an collection error in IJB-A dataset that the different persons with different genders are treated as a same person. Fig. 12(b) illustrates the top-30 worst Non-matched pairs that two different persons in the compared pairs are determined to be the same person. The ID of the two sets and the similarity score given by our approach between the two sets are shown. It can be observed that most error cases occur when the set contains only one media item, which means very little information is available for the verification.  $kNN$ -average pooling is not able to play a positive role in this case. Moreover, it is observed that the only media sample usually has an extreme pose or the face is very blurred or occluded. In the ninth line in (a), there is even an age gap between the two sets of the same person. It seems that only a noise face can be used and our method inevitably gives rise to a poor result. Most of the existing methods can not solve this issue caused



Fig. 12: Failure cases analysis. (a) Top 30 worst matched; (b) Top 30 worst Non-matched. Most face sets just contain one media and most faces are very profile.

by the very limited number of media in a set. In future, we hope to utilize synthesized images to augment data for subjects with limited samples. We assume enough samples of a subject should address the problem.

4) *Comparisons with benchmarks:* We compare our methods using different metrics with the existing 12 methods from other publications. OpenBR and GOTS are two basic baselines given in [7]. They are not deep learning methods and give very poor performances for both verification and identification. Other methods using deep convolutional networks, which were introduced in Section II, have some improvements. Not all of the methods performs on all protocols of IJB-A. For example, B-CNN method just focuses on face identification and gets the accuracy of 55.8% for rank-1, and the missed rate of 85.7% for FPIR=0.01. Most of the other methods only concentrated on the face verification task and closed-set identification. DCNNs [11] shows the best results among these benchmarks in TABLE II. However, we test all the protocols of IJB-A using  $kNN$ -average pooling based on Cosine similarity, Euclidean distance and  $JB$  metric. The results of the three settings are very competitive. Especially, the two unsupervised metrics embedded get top accuracies over most of all protocols. Even  $JB$  embedded achieve 15.1% and 32.4% respectively for FNIR@FPIR=0.1 and 0.01, which are much better than the previous results. Seen from the table,  $kNN$ -average pooling using Cosine similarity further exceeds other methods for face verification and closed-set face identification and  $kNN$ -average pooling using Euclidean distance generates large margins over other methods. In all, even though the  $kNN$ -average pooling combining with unsupervised metrics is simple, it is very effective and flexible.

## V. CONCLUSION

In this paper, we have presented a very simple but effective S2S distance to measure the similarity between two image sets, which is suitable for addressing the identification of faces with heterogeneous contents, such as the IJB-A dataset. The S2S distance is defined based on averaging a mutual comparison between the probe set and the gallery set, in which each element is only compared with its nearest neighbors. By doing so, the impact of outliers and noises commonly occurring in the IJB-A dataset is under control. Furthermore, variable metric methods can be embedded into the proposed S2S distance measure. Finally, we showed that even with a simple unsupervised metric, like Euclidean distance, the proposed method can achieve competitive accuracy for both face verification and identification. Furthermore, the  $kNN$ -average pooling is very generic and can be potentially applied to various applications over sets, where some severe noise or extreme cases occur caused by capturing conditions in real world. Such applications include person-reidentification [47], object recognition, saliency detection [48], [49] and image retrieval [50].

In future work, our system can be enhanced from two aspects: (1) we will embed the  $kNN$ -average pooling to a deep network in an end-to-end fashion. We believe such a strategy will further boost the recognition performance; (2) We will investigate the local features based on extracting key points from faces and design a S2S distance based on these local features, enabling more robust matching between faces with severe variations.

## REFERENCES

- [1] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.
- [2] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
- [3] F. Schroff, D. Kalenichenko, and J. Pjilbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [4] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Web-scale training for face identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [5] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Tech. Rep. TR-07-49, 2007.
- [6] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *International Conference on Computer Vision*, 2009.
- [7] B. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [8] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [9] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 25, no. 8, pp. 1309–1321, 2015.
- [10] J. Zabalza, J. Rena, J. Zhengb, H. Zhaoc, C. Qingd, Z. Yange, P. Duf, and S. Marshalla, "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging," *Neurocomputing*, vol. 185, pp. 1–10, 2016.
- [11] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in *IEEE WACV*, 2016.

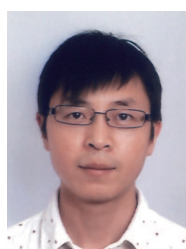
- [12] S. Sankaranarayanan, A. Alavi, and R. Chellappa, "Triplet similarity embedding for face verification," 2016.
- [13] S. Sankaranarayanan, A. Alavi, C. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," 2016.
- [14] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *British Machine Vision Conference*, 2013.
- [15] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Tang, "Person re-identification by dual-regularized kiss metric learning," *IEEE Trans. Image Processing*, vol. 25, no. 6, pp. 2726–2738, 2016.
- [16] D. Tao, J. Cheng, X. Gao, and C. Deng, "Robust sparse coding for mobile image labeling on the cloud," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 27, no. 1, pp. 62–72, 2017.
- [17] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing kiss metric learning," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 23, no. 10, pp. 1675–1685, 2013.
- [18] H. Cevikalp and B. Triggs, "Face recognition based on image set," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [19] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *International Conference on Machine Learning*, 2008.
- [20] Z. Huang, R. Wang, S. Shan, and X. Chen, "Projection metric learning on grassmann manifold with application to video based face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [21] S. Labeznik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
- [22] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [23] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014.
- [24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Image classification with deep convolutional neural networks," in *Advances in neural information processing system*, 2012, pp. 1097–1155.
- [25] A. RoyChowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller, "One-many face recognition with bilinear cnns," in *IEEE WACV*, 2016.
- [26] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan, R. Nevatia, and G. Medioni, "Face recognition using deep multi-pose representation," in *IEEE WACV*, 2016.
- [27] T. Hassner, I. Masi, J. Kim, J. Choi, and S. Harel, "Pooling faces: Template based face recognition with pooled face images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [28] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [29] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *European Conference on Computer Vision*, 2016.
- [30] D. Wang, C. Otto, and A. K. Jain, "Face search at scale: 80 million gallery," MSU, Tech. Rep. MSU-CSE-15-11, 2015.
- [31] R. Chellappa, J.-C. Chen, R. Ranjan, S. Sankaranarayanan, A. Kumar, V. M. Patel, and C. D. Castillo, "Towards the design of an end-to-end automated system for image and video-based recognition," [http://ita.ucsd.edu/workshop/16/files/paper/paper\\_2663.pdf](http://ita.ucsd.edu/workshop/16/files/paper/paper_2663.pdf).
- [32] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," 2016.
- [33] J. Ren, "Ann vs. svm: Which one performs better in classification of mccs in mammogram imaging," *Knowledge-Based Systems*, vol. 26, pp. 144–153, 2012.
- [34] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [35] R. Duda, P. Hart, and D. Stork, *Pattern classification*. Wiley, New York, 2001.
- [36] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbour classification," in *Advances in neural information processing system*, 2006, pp. 1473–1480.
- [37] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Asian Conference on Computer Vision*, 2010.
- [38] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *European Conference on Computer Vision*, 2012.
- [39] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognition*, pp. 1771–1782, 2000.
- [40] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," 2015.
- [41] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [42] J. Hu, J. Lu, and Y. P. Tan, "Discriminative deep metric learning for face verification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
- [43] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of lfw benchmark or not?" 2015.
- [44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.
- [45] H. V. Nguyen, L. Bai, and L. Shen, "Local gabor binary pattern whitened pca: A novel approach for face recognition from single image per person," in *Advances in Biometrics*, 2009, pp. 269–278.
- [46] L. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, 2008.
- [47] S. Tan, F. Zheng, L. Liu, J. Han, and L. Shao, "Dense Invariant Feature Based Support Vector Ranking for Cross-Camera Person Re-identification," *IEEE Trans. Circuits and Systems for Video Technology*, DOI: 10.1109/TCSVT.2016.2555739, 2017.
- [48] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency Detection Based on Intraliency Prior Transfer and Deep Intersaliency Mining," *IEEE Trans. Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1163–1176, 2016.
- [49] D. Zhang, D. Meng, and J. Han, "Co-Saliency Detection via a Self-Paced Multiple-Instance Learning Framework," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 865–878, 2017.
- [50] Y. Guo, G. Ding, L. Liu, J. Han, and L. Shao, "Learning to Hash With Optimized Anchor Embedding for Scalable Retrieval," *IEEE Trans. Image Processing*, vol. 26, no. 3, pp. 1344–1354, 2017.



**Jiaojiao Zhao** currently is a Ph. D. candidate with the Department of Computer Science at Northumbria University, UK.



**Jungong Han** is a faculty member with the School of Computing and Communications at Lancaster University, Lancaster, UK. Previously, he was a senior lecturer with the Department of Computer and Information Sciences at Northumbria University, UK.



**Ling Shao** (M09-SM10) is a professor with the School of Computing Sciences at the University of East Anglia, Norwich, UK. Previously, he was a professor with the Department of Computer and Information Sciences at Northumbria University, UK.